TEHCNICAL REPORT
R-381
February 2012

Controlling Selection Bias in Causal Inference

Elias Bareinboim
Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA. 90095
eb@cs.ucla.edu

Judea Pearl
Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA. 90095
judea@cs.ucla.edu

Abstract

Selection bias, caused by preferential exclusion of samples from the data, is a major obstacle to valid causal and statistical inferences; it cannot be removed by randomized experiments and can hardly be detected in either experimental or observational studies. This paper highlights several graphical and algebraic methods capable of mitigating and sometimes eliminating this bias. These nonparametric methods generalize previously reported results, and identify the type of knowledge that is needed for reasoning in the presence of selection bias. Specifically, we derive a general condition together with a procedure for deciding recoverability of the odds ratio (OR) from s-biased data. We show that recoverability is feasible if and only if our condition holds. We further offer a new method of controlling selection bias using instrumental variables that permits the recovery of other effect measures besides OR.

1 Introduction

Selection bias is induced by preferential selection of units for data analysis, usually governed by unknown factors including treatment, outcome and their consequences. Case-control studies in Epidemiology are particularly susceptible to such bias, e.g., cases may be reported only when the outcome (disease or complication) is unusual, while non-cases remain unreported (see (Glymour and Greenland, 2008; Robins et al., 2000; Robins, 2001; Hernán et al., 2004)).

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22.

To illuminate the nature of this bias, consider the model of Fig. 1 (a) in which S is a variable affected by both X (treatment) and Y (outcome), indicating entry into the data pool. Such preferential selection to the pool amounts to conditioning on S, which creates spurious association between X and Y through two mechanisms. First conditioning on S induces spurious association between its parents, X and Y. Second, S is also a descendant of a "virtual collider" Y, whose parents are X and the error term U_Y (also called "omitted factors" or "hidden variable") which is always present, though often not shown in the diagram.[1]

A medical example of selection bias was reported in (Horwitz and Feinstein, 1978), and subsequently studied in (Hernán et al., 2004; Geneletti et al., 2009), in which it was noticed that the effect of Oestrogen (X) on Endometrial Cancer (Y) was overestimated in the data studied. One of the symptoms of the use of Oestrogen is vaginal bleeding (W) (Fig. 1(c)), and the hypothesis was that women noticing bleeding are more likely to visit their doctors, causing women using Oestrogen to be overrepresented in the study.

In causal inference studies, the two most common sources of bias are confounding (Fig. 1(b)) and selection (Fig. 1(a)). The former is a result of treatment X and outcome Y being affected by a common omitted variables $\mathbf{U}$, while the latter is due to treatment or outcome (or its descendants) affecting the inclusion of the subject in the sample (indexed by S). In both cases, we have unblocked extraneous "flow" of influence between treatment and outcome, which appear under the rubric of "spurious correlation." It is called spurious because it is not part of what we seek to estimate – the causal effect of X on Y in the target population. In the case of confounding, bias occurs because we cannot condition on the unmeasured confounders, while in selection, the distribution is always conditioned on S.

[1] See (Pearl, 2009, pp. 339-341) for further explanation of this bias mechanism.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
FEB 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Controlling Selection Bias in Causal Inference	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Los Angeles,Department of Computer Science,Los Angeles,CA,90095**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Selection bias, caused by preferential exclu- sion of samples from the data, is a major obstacle to valid causal and statistical infer- ences; it cannot be removed by randomized experiments and can hardly be detected in either experimental or observational studies. This paper highlights several graphical and algebraic methods capable of mitigating and sometimes eliminating this bias. These non- parametric methods generalize previously re- ported results, and identify the type of knowl- edge that is needed for reasoning in the pres- ence of selection bias. Speci cally, we derive a general condition together with a procedure for deciding recoverability of the odds ratio (OR) from s-biased data. We show that re- coverability is feasible if and only if our condi- tion holds. We further o er a new method of controlling selection bias using instrumental variables that permits the recovery of other e ect measures besides OR.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**15**	

Formally, the distinction between these biases can be articulated thus: confounding bias is any $X - Y$ association that is attributable to selective choice of treatment, while selection bias is any association attributable to selective inclusion in the data pool. Operationally, confounding bias can be eliminated by randomization – selection bias cannot. Given this distinction, the two biases deserve different qualitative treatment and entail different properties, which we explore in this paper. Remarkably, there are special cases in which selection bias can be detected even from observations, as in the form of a non-chordal undirected component (Zhang, 2008).

As an interesting corollary of this distinction, it was shown (Pearl, 2010) that confounding bias, if such exists, can be amplified by conditioning on an instrumental variable Z (Fig. 1(d)). Selection bias, on the other hand, remains invariant under such conditioning.

We will use instrumental variables for the removal of selection bias in the presence of confounding bias, as shown in the scenario of Fig. 1(f). Whereas instrumental variables cannot ensure nonparametric identification of average causal effects, they can help provide reasonable bounds on those effects as well as point estimates in some special cases (Balke and Pearl, 1997). Since the bounding analysis assumed no selection bias, the question arises whether similar bounds can be derived in the presence of selection bias. We will show that selection bias can be removed entirely through the use of instrumental variables, therefore, the bounds on the causal effect will be narrowed to those obtained under the selection-free assumption.

This result is relevant in many areas because selection bias is pervasive in almost all empirical studies, including Machine Learning, Statistics, Social Sciences, Economics, Bioinformatics, Biostatistics, Epidemiology, Medicine, etc. For instance, one version of selection bias was studied in Economics, and led to the celebrated method developed by (Heckman, 1970). It removes the bias through a two-step process which assumes linearity, normality and, a probabilistic model of the selection mechanism.

Machine learning tasks suffer from a similar problem when training samples are selected preferentially, depending on feature-class combinations that differ from those encountered in the target environment (Zadrozny, 2004; Smith and Elkan, 2007; Storkey, 2009; Hein, 2009).

In Epidemiology, the prevailing approach is due to James Robins (Robins et al., 2000; Hernán et al., 2004), which assumes knowledge of the probability of selection given treatment. In some special cases, this probability can be estimated from data, requiring a record, for each treatment given, whether a follow up outcome (Y) is reported or not. We do not rely on such knowledge in this paper but assume, instead, that no data of treatment or outcome is available unless a case is reported (via S).

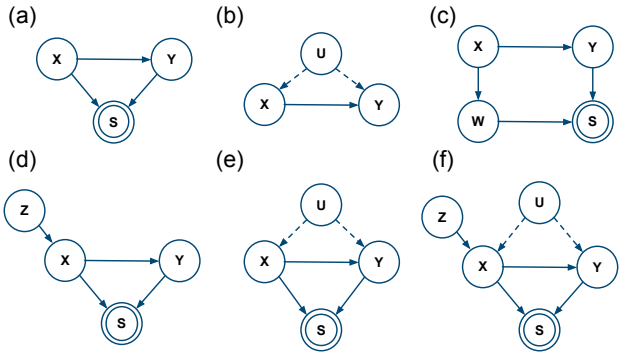


Figure 1: Different scenarios considered in this paper. (a,b) Simplest examples of selection and confounding bias, respectively. (c) Typical study with intermediary variable W between X and selection. (d) Instrumental variable with selection bias. (e) Selection combined with confounding. (f) Instrumental variable with confounding and selection bias simultaneously present.

Contributions

Our contributions are as follows. In Section 2, we give a complete graphical condition under which the population odds ratio (OR) and a covariate-specific causal odds ratio can be recovered from selection-biased data (Theorem 1). We then devise an effective procedure for testing this condition (Theorem 2, 3). These results, although motivated by causal considerations, are applicable to classification tasks as well, since the process of eliminating selection bias is separated from that of controlling for confounding bias.

In Section 3, we present universal curves that show the behavior of OR as the distribution $P(y \mid x)$ changes, and how the risk ratio (RR) and risk difference (RD) are related to OR. We further show that if one is interested in recovering RR and RD under selection bias, knowledge of $P(X)$ is sufficient for recovery.

In Section 4, we advance for other measures of effects besides odds ratio, and show that even when confounding and selection biases are simultaneously present (Fig. 1(e)), the latter can be entirely removed with the help of instrumental variables (Theorem 4). This result is surprising for two reasons: first, we generally do not expect selection bias to be removable; second, bias removal in the presence of confounding is generally expected to be a more challenging task. We finally show how this result is applicable to scenarios where other structural assumptions hold, for instance, when an instrument is not available but a certain back-door admissible set can be identified (Corollary 4).

2 Selection bias in a chain structure and its graphical generalizations

The chain structure of Figure 2(a) is the simplest structure exhibiting selection bias. The intuition gained from analyzing this example will serve as a basis for subsequently treating more complicated structures.

Consider a study of the effect of a training program (X) on earnings after 5 years of completion (Y), and assume that there is no confounding between treatment and outcome. Assume that subjects achieving higher income tend to report their status more frequently than those with lower income. The qualitative causal assumptions are depicted in Fig. 2(a). Given that all available data is obtained under selection bias, is the unbiased odds ratio recoverable?

To address this problem, we explicitly add a variable S to represent the selection mechanism, and assume that $S = 1$ represents presence in the sample, and zero otherwise. We will refer to samples selected by such mechanism as "s-biased". A similar representation was used in (Cooper, 1995; Lauritzen and Richardson, 2008; Geneletti et al., 2009; Didelez et al., 2010). In the chain structure of Fig. 2(a), X is d-separated from S by Y, which implies the conditional independence $(X \perp\!\!\!\perp S \mid Y)$, and encodes the assumption that entry to the data pool is determined by the outcome Y only, not by X. We define next some key concepts used along the paper and state some results that will support our analysis.

Definition 1 (Odds ratio)**.** *Consider two variables X and Y and a set $\mathbf{Z}$, the conditional odds ratio $OR(Y, X \mid \mathbf{Z} = \mathbf{z})$ is given by the ratio: $\big(Pr(y \mid \mathbf{z}, x')/Pr(y' \mid \mathbf{z}, x')\big)/\big(Pr(y \mid \mathbf{z}, x)/Pr(y' \mid \mathbf{z}, x)\big)$.*

$OR(Y, X \mid \mathbf{Z})$ measures the strength of association between X and Y conditioned on $\mathbf{Z}$ and it is symmetric, i.e., $OR(Y, X \mid \mathbf{Z}) = OR(X, Y \mid \mathbf{Z})$.

Definition 2 (G-Recoverability)**.** *Given a graph G, $OR(X, Y \mid \mathbf{Z})$ is said to be G-recoverable from s-biased data if the assumptions embedded in G renders it expressible in terms of the observable distribution $P(\mathbf{V_{xy}} \mid S = 1)$ where $\mathbf{V_{xy}} = \mathbf{V} \setminus \{S\}$. Formally, for every two probability distributions $P_1(.)$ and $P_2(.)$ compatible with G, $P_1(\mathbf{v_{xy}} =\mid S = 1) = P_2(\mathbf{v_{xy}} \mid S = 1)$ implies $OR_1(X, Y \mid \mathbf{Z}) = OR_2(X, Y \mid \mathbf{Z})$.*

Definition 3 (Collapsibility)**.** *Consider two variables X and Y and disjoint sets $\mathbf{Z}$ and $\mathbf{W}$. We say that the odds ratio $OR(X, Y \mid \mathbf{Z}, \mathbf{W})$ is collapsible over $\mathbf{W}$ if $OR(X, Y \mid \mathbf{Z} = \mathbf{z}, \mathbf{W} = \mathbf{w}) = OR(X, Y \mid \mathbf{Z} = \mathbf{z}, \mathbf{W} = \mathbf{w}') = OR(X, Y \mid \mathbf{Z} = \mathbf{z})$, for all $\mathbf{w} \neq \mathbf{w}'$.*

Definition 3 and the following Lemma are stated in (Didelez et al., 2010) and are based on long tradition in Epidemiology starting with (Cornfield, 1951) and followed by (Whittemore, 1978; Geng, 1992).[2]

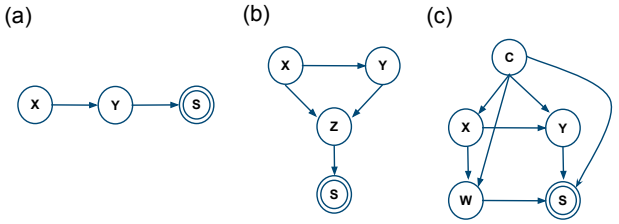


Figure 2: (a) Chain graph where X represents treatment, Y is the outcome, and S an indicator variable for the selection mechanism. (b) Scenario where there exists a blocking set from $\{X, Y\}$ to S yet the OR is not G-recoverable. (c) Example where the **c**-specific OR is G-recoverable.

Lemma 1. *For any two sets, $\mathbf{Z}$ and $\mathbf{W}$, the conditional odds ratio $OR(Y, X \mid \mathbf{Z}, \mathbf{W})$ is collapsible over $\mathbf{W}$ (that is, $OR(Y, X \mid \mathbf{Z}, \mathbf{W}) = OR(Y, X \mid \mathbf{Z})$), if either $(X \perp\!\!\!\perp \mathbf{W} \mid \{Y, \mathbf{Z}\})$ or $(Y \perp\!\!\!\perp \mathbf{W} \mid \{X, \mathbf{Z}\})$.*

The following Corollary provides a graphical test for G-recoverability (Def. 2) based on Lemma 1:

Corollary 1. *Given a graph G in which node S represents selection, the $OR(X, Y \mid \mathbf{Z})$ is G-recoverable from s-biased data if $\mathbf{Z}$ is such that $(X \perp\!\!\!\perp S \mid \{Y, \mathbf{Z}\})_G$ or $(Y \perp\!\!\!\perp S \mid \{X, \mathbf{Z}\})_G$.*

There is an important subtlety here. One might surmise that selection bias of $OR(X, Y)$ can be removed if the condition of Corollary 1 holds, i.e., there exists a separating set $\mathbf{Z}$ such that $(X \perp\!\!\!\perp S \mid \{Y, \mathbf{Z}\})_G$ or $(Y \perp\!\!\!\perp S \mid \{X, \mathbf{Z}\})_G$, but this is not the case. Consider Fig. 2(b) where the set $\mathbf{Z}$ d-separates $\{X, Y\}$ from S and therefore permits us to remove S by writing $OR(X, Y \mid \mathbf{Z}, S = 1)$ as $OR(X, Y \mid \mathbf{Z})$, yet the unconditional OR is not G-recoverable because we cannot re-apply the condition of Corollary 1 to eliminate $\mathbf{Z}$ from $OR(X, Y \mid \mathbf{Z})$. Moreover, the resulting quantity, $OR(X, Y \mid \mathbf{Z})$, though estimable for every level $\mathbf{Z} = \mathbf{z}$, does not represent a meaningful relation for decision making or interpretation, because it does not stand for a causal effect in a stable subset of individuals (see discussion about the causal OR at the end of this section). Since $\mathbf{Z}$ is X-dependent in G, the class of units for which $\mathbf{Z} = \mathbf{z}$ under $do(X = 1)$ is not the same as the class of units for which $\mathbf{Z} = \mathbf{z}$ under $do(X = 0)$. The conditional odds ratio $OR(X, Y \mid \mathbf{Z})$ would be meaningful only if $\mathbf{Z}$ is restricted to pre-treatment covariates, which are X-invariant, hence stable.

[2] Cornfield's result and some of its graphical ramifications were brought to our attention by Sander Greenland. See also (Greenland and Pearl, 2011).

We next introduce a criterion, followed by a procedure to decide whether it is legitimate to replace $\mathbf{Z}$ with a set $\mathbf{C}$ of pre-treatment covariates, for which $OR(Y, X \mid \mathbf{C})$ is a meaningful $\mathbf{c}$-specific causal effect. Typical examples of $\mathbf{c}$-specific effects would be $\mathbf{C} = \{age, sex\}$ or, when average behavior is desired, $\mathbf{C} = \{\}$.

Definition 4 (OR-admissibility)**.** *A set* $\mathbf{Z} = \{Z_1, ..., Z_n\}$ *is OR-admissible relative to an ordered triplet* $(X, Y, \mathbf{C})$ *whenever an ordering* $(Z_1, ..., Z_n)$ *exists such that for each* Z_k*, either* $(X \perp\!\!\!\perp Z_k \mid \mathbf{C}, Y, Z_1, ..., Z_{k-1})$ *or* $(Y \perp\!\!\!\perp Z_k \mid \mathbf{C}, X, Z_1, ..., Z_{k-1})$.

Corollary 2 (Didelez et al. (2010))**.** *OR-admissibility of* $\mathbf{Z}$ *implies* $OR(Y, X \mid \mathbf{C}, \mathbf{Z}) = OR(Y, X \mid \mathbf{C})$.

This Corollary follows by successive application of Lemma 1 to the elements $Z_1, ..., Z_n$ of $\mathbf{Z}$.

Theorem 1 (OR G-recoverability)**.** *Let graph* G *contain the arrow* $X \to Y$ *and a set* $\mathbf{C}$ *of measured* X*-independent covariates. The* $\mathbf{c}$*-specific odds ratio* $OR(Y, X \mid \mathbf{C})$ *is* G*-recoverable from s-biased data if and only if there exists an additional set* $\mathbf{Z}$ *of measured variables such that the following conditions hold in* G*:*

1. $(X \perp\!\!\!\perp S \mid \{Y, \mathbf{Z}, \mathbf{C}\})_G$ *or* $(Y \perp\!\!\!\perp S \mid \{X, \mathbf{Z}, \mathbf{C}\})_G$.
2. $\mathbf{Z}$ *is OR-admissible relative to* $(X, Y, \mathbf{C})$.

Moreover, $OR(Y, X \mid \mathbf{C}) = OR(Y, X \mid \mathbf{C}, \mathbf{Z}, S = 1)$. [3]

Proof. See Appendix. □

Note that unlike the control of confounding, which requires averaging over the adjusted covariates, a single instantiation of the variables in $\mathbf{Z}$ is all that is needed for removing selection bias.

Let us consider the causal story of section 1 concerning the effect of Oestrogen (X) on Endometrial Cancer (Y) as depicted in in Fig. 1(c). This problem is solvable by setting $\mathbf{Z} = \{W\}$ and applying Theorem 1 – we can readily verify that $\mathbf{Z}$ is OR-admissible relative to $(X, Y, \{\})$ (i.e., $(W \perp\!\!\!\perp Y \mid X)$), and $(X \perp\!\!\!\perp S \mid \{Y, W\})$ holds. Thus, we can write $OR(Y, X) = OR(Y, X \mid W) = OR(X, Y \mid W) = OR(X, Y \mid W, S = 1)$, which shows a mapping from the target (unbiased) quantity (without any S) to the s-biased data (conditioned on $S = 1$, which was measured). (In the sequel we will drop G finding no need to distinguish conditional independencies from d-separation statements.) [4]

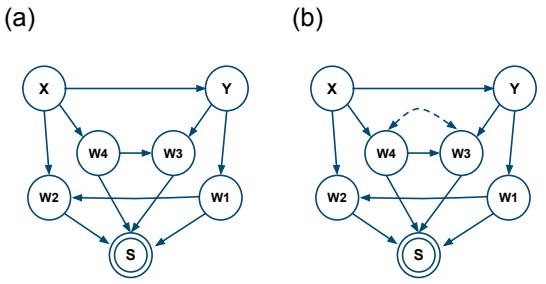


Figure 3: Scenario where OR is G-recoverable and $\mathbf{Z} = \{W_1, W_2, W_4\}$ (a), and it is not G-recoverable in (b).

Theorem 1 defines the boundary that distinguishes the class of graphs that permit G-recoverability of OR from those that do not. To show the power of Theorem 1, let us consider the more intricate scenario of Fig. 3(a), in which $\mathbf{Z} = \{W_1, W_2, W_4\}$ satisfies the conditions of Theorem 1. This can be seen through the following sequence of reductions verified by the graph: $(X \perp\!\!\!\perp S \mid \{Y, W_1, W_2, W_4\}) \to (Y \perp\!\!\!\perp W_2 \mid \{X, W_1, W_4\}) \to (X \perp\!\!\!\perp W_1 \mid \{Y, W_4\}) \to (Y \perp\!\!\!\perp W_4 \mid X)$. The final result is

$$OR(Y, X) = OR(Y, X \mid W_1, W_2, W_4, S = 1)$$

where the term on the left is our target quantity and the one on the right is estimable from the s-biased data. Fig. 3(b) shows an example where OR is not G-recoverable, because we must start with $\mathbf{Z} = \{W_1, W_2, W_3, W_4\}$ or $\mathbf{Z} = \{W_1, W_3, W_4\}$ to separate S from X or Y, respectively – these two sets are not OR-admissible since each set contains the variable W_3 which cannot be separated from X or Y by any set.

Theorem 1 relies on OR-admissibility, for which Definition 4 gives a declarative, non-procedural criterion. Taken literally, it requires that we first find a proper $\mathbf{Z}$ and then, out of the $n!$ orderings of the elements in Z, find one that will satisfy the d-separation tests specified in Definition 4. We will now supplement Theorem 1 with a simple graphical condition, followed by an effective procedure for finding such a sequence if one exists.

Theorem 2. *Let graph* G *contain the arrow* $X \to Y$*, a necessary condition for* G *to permit the* G*-recoverability of* $OR(Y, X \mid \mathbf{C})$ *for a given set* $\mathbf{C}$ *of pre-treatment covariates is that* S *and every ancestor* A_i *of* S *that is also a descendant of* X *have a separat-*

[3] This Theorem builds on and extends the results in (Didelez et al., 2010) which are summarized by Definition 4 and Corollary 2. First, it supplements the sufficient condition with its necessary counterpart. This is made possible by defining G-recoverability in terms of identifiability (Def. 2). Second, Theorem 1 explicitly avoid meaningless ORs (i.e., $OR(X, Y \mid \mathbf{Z})$, where $\mathbf{Z}$ is X-dependent). Finally, the proof of the sufficiency part prepares the ground for a procedure for finding an admissible sequence if such exists, to be shown next.

[4] Furthermore, the graph symmetric to Fig. 1(c) where the positions of X and Y are interchanged yields the same result. Similarly, another common variant of Fig. 1(c), with the edge $X \to W$ reversed, is solvable as well.

ing set $\mathbf{T_i}$ *that either d-separates* A_i *from* X *given* Y*, or d-separates* A_i *from* Y *given* X*.* [5]

Proof. See Appendix. □

Theorem 3. *Let* G *be a DAG containing the arrow* $X \to Y$ *and two sets of variables, measured* $\mathbf{V}$ *and unmeasured* $\mathbf{U}$*. A necessary and sufficient condition for* G *to permit the* G*-recoverability of* $OR(Y, X \mid \mathbf{C})$ *for a given set* $\mathbf{C}$ *of pre-treatment variables is when the sink-procedure below terminates. Moreover,* $OR(Y, X \mid \mathbf{C}) = OR(Y, X \mid \mathbf{C}, \mathbf{Z}, \mathbf{T}, S = 1)$*, where* $\mathbf{Z} = \big(An(S) \setminus An(Y)\big) \cap \mathbf{V}$ *and* $\mathbf{T}$ *is given by the sink-procedure.*

Procedure (Sink reduction)

1. Set $\mathbf{T} = \{\}$, and consider $\mathbf{Z}$ as previously defined. Remove $\mathbf{V} \setminus An(Y \cup S)$ from $\mathbf{G}$, and name the new graph $\mathbf{G}^*$. Consider an ordering compatible with $\mathbf{G}^*$ such that $Z_i < Z_j$ whenever Z_i is non-descendant of Z_j.

2. Test if sink Z_i of $\mathbf{G}^*$ satisfies the following condition: $(Z_i \perp\!\!\!\perp X \mid C, T, Y, Z_1, ..., Z_i - 1)$ or $(Z_i \perp\!\!\!\perp Y \mid C, T, X, Z_1, ..., Z_i - 1)$. If so, go to step 4. Otherwise, continue.

3. Test if there exists a minimal set $\mathbf{T_i}$ of non-descendants of $\mathbf{X}$ that, if added to $\mathbf{T}$ would render step 2 successful, if none exists, exit with failure.[5] Else, add $\mathbf{T_i}$ to $\mathbf{T}$ and continue with step 4.

4. Remove Z_i from $\mathbf{G}^*$ and $\mathbf{Z}$, and repeat step 2 recursively until $\mathbf{Z}$ is empty. If so, go to step 5.

5. Test if $(\mathbf{T} \perp\!\!\!\perp Y \mid \mathbf{C}, X)$, if so, the sequence $(Z_1, Z_2, ...Z_m)$ with $\mathbf{T}$ constitutes a witness for the OR-admissibility of $\mathbf{Z}$ relative to $(X, Y, \mathbf{C})$, for a set $\mathbf{C}$ of X-independent variables. Otherwise, exit with failure.

Proof. See Appendix. □

The algorithm exploits the graph structure to construct a mapping from the observed s-biased data and the desired target OR. Since the OR is symmetric, it is not necessary to separate S from X and Y simultaneously, but only from one of them (given the other.) For simplicity, denote the expression "X given Y or Y given X" by the symbol Φ_{xy}. A separating set from S to Φ_{xy} is first sought in step 2, starting with all observable ancestors of S that are non-ancestors of Y. If the test succeeds and this set is a separator, the algorithm iterates trying to separate Φ_{xy} from the deepest node in the remaining set. In case of failure, the algorithm attempts (step 3) to achieve separability using pre-treatment covariates $\mathbf{T_i}$. In case no separability can be found using these added covariates, the algorithm fails. Otherwise, at the end, the algorithm further requires that all $\mathbf{T_i}$ added along these iterations be separable from Y (step 5).

To illustrate, running the procedure on the graph of Fig. 3(b) with $\mathbf{C} = \{\}$, the graph remaining after the removal of S has two sink nodes, W_2 and W_3. Removing W_2 leaves two other sinks, W_3, and W_1. Removing W_1 leaves W_3 as the only remaining sink node which fails the test of Step 3. Since no non-descendant of X exists that yields separability, we must exit with failure. On the other hand, if we are able to measure U, the hidden variable responsible for the double arrow arc between W_3 and W_4, we would add this node to $\mathbf{T}$, W_3 will pass the test, followed by W_4, and we will end up with U as the only non-descendant of X remaining in $\mathbf{T}$. In step 5 we remove U from $\mathbf{T}$, yielding $OR(X, Y) = OR(X, Y \mid \mathbf{W}, U, S = 1)$.

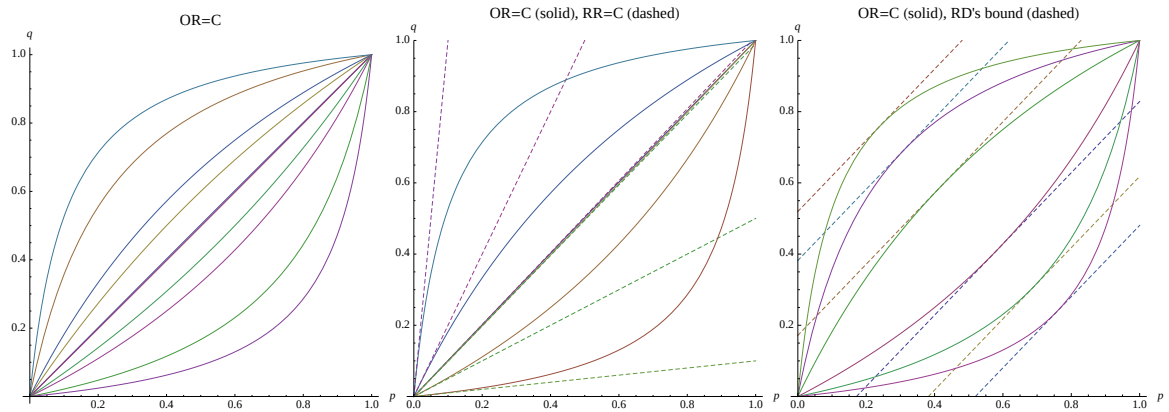
Thus far, we assumed that the treatment X is unconfounded, therefore the OR is identical to the causal OR defined as $COR(X, Y) = \frac{P(y|do(x))P(y'|do(x'))}{P(y|do(x'))P(y'|do(x))}$. In the presence of confounding, it is not enough to recover OR in s-biased data, we need to go further and assure that the recovered $OR(X, Y \mid \mathbf{C})$ is such that C satisfies the back-door criterion (2nd rule of do-calculus, observing and intervening are equivalent), in which case $OR(X, Y \mid \mathbf{C})$ will represent the **c**-specific causal OR. For example, in Fig. 2(c) the $COR(X, Y \mid \mathbf{C})$ will be G-recoverable because once we condition on $\mathbf{C}$ all conditional independencies will be identical to those of Fig. 1(c), and $P(Y \mid do(X), \mathbf{C}) = P(Y \mid X, \mathbf{C})$.

Note, however that although we can recover the **c**-specific causal OR, we cannot recover the population $COR(X, Y)$. For such measure to be recoverable we need to add assumptions which will make it possible to infer averageable measures of causal effects such as RD and RR, to be handle next.

3 OR and other measures of causal effects

Consider again the chain structure in Fig. 2(a) and define the causal effect as $COR(X, Y)$. The fact that X and Y are not confounded permits us to estimate the causal effect $COR(X, Y)$ by the odd ratio $OR(X, Y)$ which, by the results in the previous section, will re-

[5] A polynomial time algorithm for finding a minimal separating set in DAGs is given in (Tian et al., 1998). The *restricted minimal separation* version of that algorithm finds a minimal separator in a DAG with latent variables (equivalently, semi-Markovian models). A fast test for the non-separability of X and A_i is the existence of an inducing path between the two variables (Verma and Pearl, 1990). For example, the path $X \to W_4 \to W_3$ in Fig. 3(b).



(a) (b) (c)

Figure 4: (a) Constant odds ratio curves for $c = \{1.00, 1.01, 1.50, 2.00, 5.00, 10.00\}$ and their inverses; Superimposed constant odds ratio with constant risk ratio curves (b) and constant risk difference curves (c).

main invariant to conditioning on $S = 1$. However, if we define the causal effect as $ACE = Pr(y \mid do(x)) - Pr(y \mid do(x'))$ (also known as the causal risk difference), a bias will be introduced upon conditioning.

The invariance of OR can be represented in the following intuitive and pictorial way. We characterize the conditional distribution $P(Y \mid X)$ by two independent parameters $p = P(y \mid x)$ and $q = P(y \mid x')$, which define a point (p, q) in the unit square. The condition $OR(X, Y) = c$ describes a curve in the $(p, q)-$plane. For $c = 1$, the curve is the unit slope line. For $c > 1$, this curve separate points with $OR(.) > c$ from those with $OR(.) < c$ in the region below the unit slope line (symmetrically for the inverses ($c < 1$) in the region above $q = p$). See Fig. 4.

Now, by conditioning on $S = 1$, we obtain a new conditional probability, also characterized by two independent parameters $p_s = P(y \mid x, S = 1), q_s = P(y \mid x', S = 1)$. The fact that $OR(Y, X \mid S = 1) = OR(Y, X)$ means that conditioning on $S = 1$ must shift the initial (p, q) point along a constant OR curve, not anywhere else. We show these universal curves of constant OR for $c = \{1.00, 1.01, 1.50, 2.00, 5.00, 10.00\}$ and their respective inverses in Fig. 4(a). Fig. 4(b) shows curves for constant risk ratio (RR: $\frac{p}{q} = c$), which are variable slope lines going through the origin, and bounded by the slope $\frac{1}{c}$. Similarly, Fig. 4(c) shows curves for constant risk difference.

We see that even though RR does not remain constant (upon conditioning), the constancy of OR constrains the behavior of the RR. This follows by noting (after some algebra) that $RR = c + (1 - c)p$, i.e., RR has intercept c and slope $1 - c$. For instance, if OR is constant and $c = 1$, we have unit slope line for OR, but RR does not move and is equal to one. For constant OR and $\frac{1}{2} < c < 1$, the slope is positive but less than $\frac{1}{2}$, and the intercept is greater than $c = \frac{1}{2}$, which implies that RR lies inside the interval $[c, 1]$. Similar bounds can be obtained for other values of c.

Recovering RR and RD under selection bias

In this section we show that, in some situations, point estimates of RR and RD can be recoverable from s-biased data in studies where the prior probability $P(X)$ is available. [6] In other words, we refer back to the chain structure of Fig. 2(a) and ask whether $P(Y \mid X)$ can be recovered from $P(X)$ and $P(X, Y \mid S = 1)$.

The solution can be obtained algebraically, noting that Y d-separates X from S, which permits us to write:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{\Big(P(Y \mid X)P(X) + P(Y \mid \neg X)P(\neg X)\Big)}$$

$$P(X \mid \neg Y) = \frac{P(\neg Y \mid X)P(X)}{\Big(P(\neg Y \mid X)P(X) + P(\neg Y \mid \neg X)P(\neg X)\Big)}$$

This can be turn into a two linear equations with two unknowns, $\{P(Y \mid X), P(Y \mid \neg X)\}$, which gives:

$$P(Y \mid X) = -\frac{P(X \mid Y)\Big(P(X \mid Y) - P(X)\Big)}{\Big(P(X \mid Y) - P(X \mid \neg Y)\Big)P(X)}$$

[6] Potentially, we are under a RCT setup or have an alternative way to access it through external studies as census' data.

$$P(Y \mid \neg X) = \frac{P(\neg X \mid Y)\Big(P(X \mid \neg Y) - P(X)\Big)}{\Big(P(X \mid \neg Y) - P(X \mid Y)\Big)P(\neg X)}, \tag{1}$$

where $P(X \mid Y) = P(X \mid Y, S = 1), \forall X, Y$.[7]

This simple result exemplifies a general theme of correcting for selection bias (section 4); the bias induced by preferential selection can be removed if we have enough unconfounded variables that constraint the distribution of the remaining variables in a specific way.

Note that this case is different than as previously discussed in which we were just interested in the OR. Next we extend this result for more elaborated scenarios.

4 Randomization with non-compliance under selection bias

Let us consider the more general problem depicted in Fig. 5(a) in which confounding and selection biases are simultaneously present, and there are instrumental variables available.

Our goal is to infer the most accurate bounds for the causal effect of X on Y, knowing that there is no unbiased estimate for this quantity even when selection bias is not present. This scenario is usually presented under the rubric of "randomization with non-compliance", and it is pervasive in the Economics literature, we defer to (Pearl, 2009, Ch. 8) for a more comprehensive discussion of the relevance of this setup, we focus here on the technical aspects of the problem.

Generally, the bounding analysis assumes no selection bias, and the natural question that arises is whether selection bias can be treated and under which conditions bounds free from selection can be recovered.

We show next that this problem can be solved assuming the existence of two instrumental variables Z_1 and Z_2. [8] Noteworthy, the set of assumptions used in our analysis are commonplace in daily Econometrics practice, and its convoluted appearance is diluted when one observes them more vividly through the causal graph depicted in Fig. 5(a). In a nutshell, they are the same

[7] In Epidemiology, there are many "longitudinal data settings" where selection bias is sequential, in which it can be possible easier to estimate the probability of selection instead of $P(X)$ – this observation was brought to our attention by Onyebuchi A. Arah.

[8] Call $\mathbf{Z} = Z_1 \cup Z_2$, or consider one IV with the same number of levels. Let us name both cases by instrumental variable set.

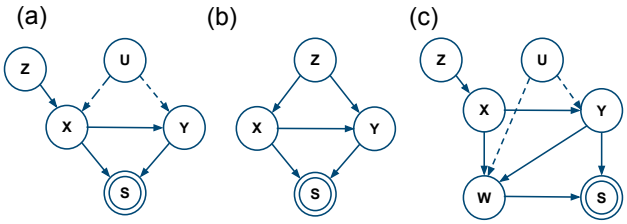


Figure 5: Different scenarios in which Theorem 4 can be applied. (a) Typical study with randomization and non-compliance (IV as incentive-mechanism) where selection and confounding are both present. (b) Selection bias in the back-door case. (c) More complex study with an intermediary variable W between treatment and selection. In this case, Y directly cause W and there is a common cause between them (extension of Fig. 1(c), see corollary 5.)

assumptions of randomization with non-compliance together with selection bias (such that treatment and outcome affect entry in the data pool).

Theorem 4. *The joint distribution of $P(X, Y, \mathbf{Z})$ is recoverable from s-biased data whenever the following conditions hold: (i) the S node is affected by the set $\mathbf{Z}$ only through $\{X, Y\}$; (ii) the set $\mathbf{Z}$ is d-connected to $\{X, Y\}$ (and combinations); (iii) the dimensionality of $\mathbf{Z}$ matches the dimensionality of $\{X, Y\}$; (iv) the marginal probability of $\mathbf{Z}$ is known. In other words, the distribution $P(X, Y, \mathbf{Z})$ is recoverable from s-biased data whenever $(S \perp\!\!\!\perp \mathbf{Z} \mid X, Y)$, $(\mathbf{Z} \not\perp\!\!\!\perp \{X, Y\})$, $(\mathbf{Z} \not\perp\!\!\!\perp X \mid Y)$,$(\mathbf{Z} \not\perp\!\!\!\perp Y \mid X)$, the dimensionality of $\mathbf{Z}$ and $X \cup Y$ matches, and the marginal distribution of $P(\mathbf{Z})$ is given.*

Proof. See Appendix. □

Corollary 3. *The bounds for $P(y \mid do(x))$ in the scenario of randomization with non-compliance (Fig. 5(a)) are recoverable from s-biased data whenever the conditions of the Theorem 4 hold.*

Proof. It follows directly from Theorem 4 together with the bounds in (Balke and Pearl, 1997). □

Corollary 4. *The causal effect $P(y \mid do(x))$ in the back-door scenario (Fig. 5(b)) is recoverable from s-biased data whenever the conditions of the Theorem 4 hold.*

Proof. It follows directly from Theorem 4. □

Corollary 5. *The causal effect of Oestrogen (X) on Endometrial Cancer (Y) as studied in (Horwitz and Feinstein, 1978; Hernán et al., 2004) (Fig. 5(c)) is recoverable from s-biased data whenever there is an IV set $\mathbf{Z}$ pointing to X, and the conditions of the Theorem*

4 hold. Moreover, the same holds without relying on $\mathbf{Z}$ *whenever the following conditions hold: (i)* X *has the same dimensionality of* $\{W, Y\}$*; (ii) the marginal distribution of* $P(X)$ *is available.*

Proof. See Appendix. □

Some observations on the method

Methods that handle selection bias under different causal assumptions try to model the distribution of S, which is unobservable and usually hard to estimate; we take a different approach and avoid doing this explicit manipulation of the selection mechanism by exploiting the topology of the causal graph and the underlying data-generating process. We are not aware of other approaches trying to do so.

The main idea is to exploit the conditional independence of the IV set $\mathbf{Z}$ and the selection mechanism S given the distribution of the treatment and outcome – interestingly, the latter is what we seek to estimate. The method hinges on two properties about the induced system, that it is linearizable and full rank – both facts were not obvious nor expected a priori.

It is worth to make some additional remarks that follow the proof of Theorem 4. First note that the proposed method relies on a sample size approaching infinity, which is difficult to obtain in practice. As a possible improvement, the problem could be cast as an optimization problem. The formulation goes as follows. We associate error terms $\epsilon_{z_1 z_2, xy}$ to each $\gamma_{z_1 z_2, xy}$ term, and proceed the analysis minimizing the (square) mean error subject to constraints. The constraints emerge naturally from the induced system of equations together with the additional constraints of positivity and integrality. Our original goal was to show feasibility of removing selection bias (identifiability) but not the estimation per se, still, this should be an interesting exercise to pursue. Further investigation is needed to check the applicability of this suggestion.

We envision our method being used as a first step in a pre-processing stage, before the application of any bounding (Balke and Pearl, 1997) or estimation procedure. The method returns the same values of $P(X, Y, \mathbf{Z})$ whenever the collected data is not under selection bias, which means that its usage will not hurt and should be considered as a "good practice."

Finally, it is also important to mention that there are scenarios not solvable by our method or in which our assumptions are not applicable. For instance, we show in Fig. 6 one of this kind, in which selection and confounding biases are entangled in such way that it does not seem possible to detach one from another. We conjecture that this case is not solvable in general without further assumptions. Notice that even if we remove the edge $U \to X$, the example is still hard to resolve.

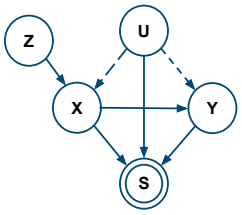


Figure 6: Scenario in which selection and confounding biases are present, entangled, and thus not recoverable.

5 Conclusion

We showed that qualitative knowledge of the selection mechanism and the use of instrumental variables can eliminate selection bias in many realistic problems. In particular, the paper provides a general graphical condition together with an algorithm that operates on a general DAG, with measured and unmeasured nodes, and decides whether and how a given **c**-specific odds ratio can be recovered from selection-biased data characterized by a selection node S. We further showed by algebraic methods that selection bias can be removed with the help of instrumental variables under a mild set of conditions.

This paper complements recent work on transportability (Pearl and Bareinboim, 2011) which deals with transferring causal information from one environment to another, in which only passive observations can be collected. The solution to the transportability problem assumes that disparities between the two environments are represented graphically in the form of unobserved factors capable of causing such disparities. The problem of selection bias also seeks extrapolation between two environments; from one in which samples are selected preferentially, to one in which no preferential sampling takes place. Both problems represent environmental differences in the form of auxiliary (selection) variables, the influence of which we seek to eliminate. However the semantics of those variables is different. In selection bias the auxiliary s-variables represent disparities in the data-gathering process, whereas in transportability problem they represent disparities in the structure of the data-generation process itself.

Acknowledgement

The authors would like to thank the reviewers for their comments that help improve the manuscript. This paper also benefited from discussions with Humberto Silva Naves, Onyebuchi Arah, and Sander Greenland. This research was supported in parts by NIH #1R01 LM009961-01, NSF #IIS-0914211, #IIS-1018922, ONR #N000-14-09-1-0665, and #N00014-10-1-0933.

Appendix – Proofs

Theorem 1

(if part) Our target quantity is $OR(X, Y \mid \mathbf{C})$ and given that $\mathbf{Z}$ is OR-admissible relative to $(X, Y, \mathbf{C})$, Corollary 2 permits us to add $\mathbf{Z}$ and rewrite it as $OR(X, Y \mid \mathbf{C}, \mathbf{Z})$. Given that the first condition of the theorem holds, Corollary 1 implies $OR(X, Y \mid \mathbf{C}, \mathbf{Z}) = OR(X, Y \mid \mathbf{C}, \mathbf{Z}, S = 1)$. This establishes G-recoverability since the r.h.s. is estimable from the available s-biased data.

(only if part) If the conditions of the theorem cannot be satisfied, then $OR(X, Y \mid \mathbf{C})$ is not G-recoverable, that is, there exist two distributions P_1, P_2 compatible with G such that they agree in the probability under selection, $P_1(\mathbf{V} \setminus \{S\} \mid S = 1) = P_2(\mathbf{V} \setminus \{S\} \mid S = 1)$, and disagree in the odds ratio, $OR_1(X, Y \mid \mathbf{C}) \neq OR_2(X, Y \mid \mathbf{C})$. We first consider the case when $\mathbf{C} = \{\}$, and we will construct two such distributions. Let P_1 be compatible with the graph $G_1 = G$, and P_2 with the subgraph G_2 where all edges pointing to S are removed. Both are compatible with G, since compatibility with a subgraph assures compatibility with the graph itself (Pearl, 1988). Notice that P_2 harbors an additional independence $(\mathbf{V} \setminus \{S\} \perp\!\!\!\perp S)_{P_2}$. By construction $P_1(X, Y \mid S = 1) = P_2(X, Y \mid S = 1)$, but since

$$P_2(X, Y|S = 1) = P_2(X, Y),$$

we have:

$$P_1(X, Y|S = 1) = P_2(X, Y)$$

We can then simplify OR_2 rewriting it as follows

$$OR_2 = \frac{P_1(X, Y, S = 1)P_1(\overline{X}, \overline{Y}, S = 1)}{P_1(\overline{X}, Y, S = 1)P_1(X, \overline{Y}, S = 1)}, \tag{2}$$

and similarly for OR_1,

$$OR_1 = \frac{P_1(X, Y)P_1(\overline{X}, \overline{Y})}{P_1(\overline{X}, Y)P_1(X, \overline{Y})} \tag{3}$$

We want to show that it is possible to produce a parametrization of P_1 in such way that $OR_1(X, Y) \neq OR_2(X, Y)$. First, let us consider the class of Markovian models. Accordingly, P_1 can be parametrized through its factors in the Markov decomposition $P_1(S = 1 \mid \mathbf{PA_s}), P_1(X \mid \mathbf{PA_x}), \ldots$, or more generally, $P_1(V_i \mid \mathbf{PA_i})$ for each family in the graph. This choice of parameters induces a valid parameterization for P_2 as well. Firstly, let us consider the case in which condition 1 of the theorem fails, i.e., $\{X, Y\}$ are not separable from S. Thus, eq. (2) can be rewritten using the identity $P_1(X, Y, S = 1) = P_1(S = 1 \mid X, Y)P_1(X, Y)$, yielding:

$$OR_2 = OR_1\left(\frac{P_1(S = 1|X, Y)P_1(S = 1 \mid \overline{X}, \overline{Y})}{P_1(S = 1 \mid \overline{X}, Y)P_1(S = 1 \mid X, \overline{Y})}\right) \tag{4}$$

Note that making the multiplier of OR_1 in eq. (4) different than 1 entails $OR_2 \neq OR_1$, which will happen for *almost all* parametrizations of $P_1(S = 1 \mid .)$ independently of the one chosen for $P_1(X, Y)$. In case there are additional nodes pointing to S, we can just make them independent of S in this new parametrization given that compatibility with the subgraph is enough to ensure compatibility with G.

Now, let us consider the case in which condition 2 of the theorem fails, i.e., there is no OR-admissible sequence in relation to $(X, Y, \{\})$. Let $\mathbf{Z} = \mathbf{V} \setminus \{X, Y, S\}$, and expand $P_1(X, Y, S = 1)$ in the following way[9]:

$$\begin{aligned} P_1(X, Y, S = 1) &= \sum_Z P_1(X, Y, S = 1, \mathbf{Z}) \\ &= \sum_Z P_1(X \mid \mathbf{PA_x}) \ldots P_1(S = 1 \mid \mathbf{PA_s}) \\ &= \sum_Z \prod_{\mathbf{V} \cap S = 1} P_1(V_i \mid \mathbf{PA_i}) \end{aligned} \tag{5}$$

Notice that each term in eq. (2) can be rearranged for each assignment of S' parents (i.e., $\mathbf{PA_s} = \mathbf{pa_s^{(j)}}$), for instance, we can write based on eq. (5):

$$\begin{aligned} &P_1(X, Y, S = 1) = \\ &P_1(S = 1 \mid \mathbf{PA_s} = \mathbf{pa_s^{(1)}}, \lambda)\Big(\sum_{\mathbf{Z}, \mathbf{PA_s} = \mathbf{pa_s^{(1)}}} \prod_{\mathbf{V} \setminus S} P_1(V_i \mid \mathbf{PA_i})\Big) + \\ &P_1(S = 1 \mid \mathbf{PA_s} = \mathbf{pa_s^{(2)}}, \lambda)\Big(\sum_{\mathbf{Z}, \mathbf{PA_s} = \mathbf{pa_s^{(2)}}} \prod_{\mathbf{V} \setminus S} P_1(V_i \mid \mathbf{PA_i})\Big) + \\ &\ldots \\ &P_1(S = 1 \mid \mathbf{PA_s} = \mathbf{pa_s^{(k)}}, \lambda)\Big(\sum_{\mathbf{Z}, \mathbf{PA_s} = \mathbf{pa_s^{(k)}}} \prod_{\mathbf{V} \setminus S} P_1(V_i \mid \mathbf{PA_i})\Big) \end{aligned} \tag{6}$$

where k is the number of configurations of S' parents, and λ indexes configurations of X or Y whenever one of them is a parent of S. Given eq. (6), let us call $P_1(S = 1 \mid \mathbf{PA_s} = \mathbf{pa_s^{(1)}}, \lambda) = \alpha_1^\lambda$, $P_1(S = 1 \mid \mathbf{PA_s} = \mathbf{pa_s^{(2)}}, \lambda) = \alpha_2^\lambda, \ldots$, and also call $\sum_{\mathbf{Z}, \mathbf{PA_s} = \mathbf{pa_s^{(j)}}} \prod_{\mathbf{V} \setminus S} P_1(V_i \mid \mathbf{PA_i}) = f_j(x, y)$ for each configuration $X = x, Y = y, \mathbf{PA_s} = \mathbf{pa_s^{(j)}}$. Then, we can write eq. (6) in the following simplified manner:

$$P_1(X, Y, S = 1) = \alpha_1^\lambda f_1(x, y) + \alpha_2^\lambda f_2(x, y) + \ldots \tag{7}$$

for all values of X and Y. We can then rewrite OR_2

[9] It clear that we should consider in the expression above (in respect to $\mathbf{Z}$) just the nodes that are somehow related to S, i.e., its ancestors, otherwise we could just sum these vertices out because they do not offer any additional constraint over the distribution of interest related to OR, and then in its respective parameterization.

based on eq. (7) as

$$OR_2 = \frac{(\alpha_1^{\lambda} f_1(x,y) + \alpha_2^{\lambda} f_2(x,y) + \ldots)}{(\alpha_1^{\overline{\lambda}} f_1(\overline{x},y) + \alpha_2^{\overline{\lambda}} f_2(\overline{x},y) + \ldots)} \times \frac{(\alpha_1^{\overline{\lambda}} f_1(\overline{x},\overline{y}) + \alpha_2^{\overline{\lambda}} f_2(\overline{x},\overline{y}) +)}{(\alpha_1^{\lambda} f_1(x,\overline{y}) + \alpha_2^{\lambda} f_2(x,\overline{y}) + \ldots)} \quad (8)$$

and similarly for OR_1:

$$OR_1 = \frac{(f_1(x,y)) + f_2(x,y) + \ldots)(f_1(\overline{x},\overline{y}) + f_2(\overline{x},\overline{y}) + \ldots)}{(f_1(\overline{x},y)) + f_2(\overline{x},y) + \ldots)(f_1(x,\overline{y}) + f_2(x,\overline{y}) + \ldots)} \quad (9)$$

There is an important observation here. Given that there is no admissible sequence relative to $(X, Y, \{\})$, there exists a set $\mathbf{W}$ such that $\mathbf{W}$ is needed to separate S from X or Y, but also $(\mathbf{W} \not\perp\!\!\!\perp \{X, Y\} \mid \mathbf{Z}')$, for $\mathbf{Z}'$ non-descendents of $\mathbf{W}$ and in $Anc(S)$, otherwise there will exist an admissible sequence. If $\mathbf{W}$ is different than $\{S\}$, it is the case that, by construction, $\mathbf{W}$ is contained in the factor $f_i(x, y)$. Thus, we have an asymmetry given that $\mathbf{W}$, and so $f_i()$, change depending *simultaneously* on the specific instantiation of X and Y, and consequently eq. (8) cannot be simplified in the general case. I.e., the linear combinations encoded in $f_i()$'s at eq. (8) do not deteriorate, factoring out independently of the given parametrization given that there is a different element in each one of them.

Now let us consider the following parametrization for P_1: set $P_1(V_i \mid \mathbf{PA_i}) = 1/2$ for all families except for the family of the S node (i.e., $P(S = 1 \mid \mathbf{PA_s})$) and the exclusive families included in the factor $f_i(x, y)$ (i.e., for when $X = x, Y = y$). Thus, rewrite OR_2 based on eq. (8):

$$OR_2 = \frac{(\alpha_1^{\lambda} f_1(x,y) + \alpha_2^{\lambda} f_2(x,y) + \ldots)}{(1/2)^l(\alpha_1^{\lambda} + \alpha_2^{\lambda} + \ldots)} \quad (10)$$

where l is equal to k minus the number of summands in the respective expression (eq. (6)). Let us also rewrite eq. (9) accordingly with this given parametrization, which yields:

$$OR_1 = \frac{(f_1(x,y) + f_2(x,y) + \ldots)}{k(1/2)^l} \quad (11)$$

After applying some simplifications on eqs. (10) and (11), we obtain, respectively,

$$OR_2 = \frac{(\alpha_1^{\lambda} f_1(x,y) + \alpha_2^{\lambda} f_2(x,y) + \ldots)}{(\alpha_1^{\lambda} + \alpha_2^{\lambda} + \ldots)} \quad (12)$$

and

$$OR_1 = \frac{(f_1(x,y) + f_2(x,y) + \ldots)}{k} \quad (13)$$

Notice that OR_2 in eq. (12) is the weighted arithmetic mean of $f_i(.)$'s averaged by α_i^{λ}'s, and OR_1 in eq. (13) is the arithmetic mean of $f_i(.)$'s. After simplifications, the remaining parameters lie in the space $[0, 1]^{m+k}$, where m is the number of free parameters in $f_i(.)$'s. Note that $OR_1 - OR_2 = 0$ adds a constraint in this space, and in order to satisfy it we should choose any point in a surface in $[0, 1]^{m+k-1}$ inside $[0, 1]^{m+k}$, i.e., which has Lebesgue measure zero. Consequently, if we randomly choose parameters the equality will *almost never* hold (and the inequality $OR_1 \neq OR_2$ *almost always*), and then just randomly draw the parameters from $[0, 1]^{m+k}$ until this is the case, which finishes this part of the proof. The case of the conditional OR is similar, and we basically have to write appropriately eqs. (2) and (3) considering $\mathbf{C}$, and exactly the same reasoning applies.

For the case when the graph contains unobservable variables, the proof is essentially the same except that an appropriate parametrization of the underlying generating model should be used – for such, consider the factorization given in (Evans and Richardson, 2011).

Theorem 2

For the necessity of the condition, we need to show that the failure of any ancestor A_i of S that is also a descendant of X (including S itself) to be separated (from either X or Y) prevents recoverability of $OR(Y, X \mid \mathbf{C})$. Indeed, A_i cannot be part of admissible sequence nor can any of its children be part of an admissible sequence, because in order to separate any such child from either X or Y we would need to condition on the father A_i, and then, the sequence will become non-admissible. Proceeding by induction, we eventually reach S itself, whose failure to enter an admissible sequence renders the existence of such sequence impossible. By Theorem 1, the inexistence of admissible sequence implies the not G-recoverability of $OR(X, Y, \mathbf{C})$. □

Theorem 3

We use along the proof some graphoid axioms and other DAG properties as shown in (Pearl, 1988). Let us first consider the correctness of the algorithm. The main idea of the reduction sequence is to use each conditional independence (CI) in step 2 of the sink-procedure to substantiate an OR reduction, creating a mapping starting from the s-biased data $OR(X, Y \mid \mathbf{C}, Z_1, ..., Z_k, S = 1)$ and reaching the target (unbiased) expression $OR(X, Y \mid \mathbf{C})$. If nodes are not added in step 3 of the algorithm, it is obvious that the sequence induces a valid step-OR reduction, which witnesses the OR G-recoverability. So, let us con-

sider the case when nodes have to be added to $\mathbf{T}$ along the execution of the algorithm. At each step i, we reduce $OR(X, Y \mid C, \mathbf{T}, Z_1, ..., Z_i)$ to $OR(X, Y \mid \mathbf{C}, \mathbf{T}, Z_1, ..., Z_{i-1})$ allowed by the CI in step 2. But given that $\mathbf{T_i}$ can be added to $\mathbf{T}$ along the execution of the algorithm, we need to show that this operation is allowed, i.e., it does not invalidate the construction of the desired mapping between the unbiased OR and the s-biased one. Towards contradiction, consider an arbitrary node Z_j such that

$$(Z_j \perp\!\!\!\perp X \mid \mathbf{C}, \mathbf{T}, Y, Z_1, ..., Z_{j-1}) \text{ or } (Z_j \perp\!\!\!\perp Y \mid \mathbf{C}, \mathbf{T}, X, Z_1, ..., Z_{j-1}) \quad (14)$$

Now, consider the first Z_k such that $k < j$ and, in order to satisfy step 2 in the sink-procedure, $\mathbf{W}$ has to be added to the conditioning set, then

$$(Z_k \perp\!\!\!\perp X \mid \mathbf{C}, \mathbf{T}, Y, Z_1, ..., Z_{k-1}, W) \text{ or } (Z_k \perp\!\!\!\perp Y \mid \mathbf{C}, \mathbf{T}, X, Z_1, ..., Z_{k-1}, W) \quad (15)$$

but also

$$(Z_j \perp\!\!\!\perp X \mid \mathbf{C}, \mathbf{T}, Y, Z_1, ..., Z_{j-1}, W) \text{ or } (Z_j \perp\!\!\!\perp Y \mid \mathbf{C}, \mathbf{T}, X, Z_1, ..., Z_{j-1}, W) \quad (16)$$

is false. If the sink-procedure ends, it is also true that

$$(\mathbf{T} \perp\!\!\!\perp Y \mid \mathbf{C}, X) \quad (17)$$

From eq. (14), all paths from Z_j to X or Y (including the ones passing through $\mathbf{W}$) are closed after conditioning on $\{\mathbf{C}, \mathbf{T}, Y, Z_1, ..., Z_{j-1}\}$. From eq. (15) and the minimal choice of T_i in step 3, it must be the case that there is a path p from Z_k to X or Y such that p is blocked by some $W \in \mathbf{W}$. From eq. (16), there exists a path p' that has to be open after condition on $\mathbf{W}$, and therefore there exists a collider U such that $U = W$ or $W \in Desc(U)$. Let us consider two possible scenarios for p', the first when it goes from Z_j to Y, and the second when it goes from Z_j to X. In the former case, there is an open path from W to Y, which is a contradiction with eq. (17) given that $\mathbf{W} \subseteq \mathbf{T}$. Then it must be the case that $\mathbf{W}$ only blocks paths ending in X, so let us assume the case in which the end node in p' is X. From (15), p is such that $Z_k \leftarrow\!\rightarrow\!\leftarrow ... - W - ... \rightarrow\!\leftarrow\!\rightarrow X$, where we are condition on all intermediate converging arrows and W must be a chain or a common cause (i.e., $\rightarrow W \rightarrow$ or $\leftarrow W \rightarrow$). Split p into $p_1 : Z_k \dots W$, and $p_2 : W \dots X$. From eq. (16), p' is such that $\mathbf{W}$ opens a collider U, then the path from Z_j to X. Split p' into $p'_1 : Z_j ... \rightarrow U$ and $p'_2 : U \leftarrow ...X$. Now we have two possibilities. If p_2 is such that $W \rightarrow \dots X$, we can concatenate $Z_k \overset{p'_1}{\rightarrow} U \rightarrow W \overset{p_2}{\rightarrow} X$, which shows an open path from Z_k to X even before conditioning on W, contradiction. If p_2 is such that $W \leftarrow \dots X$, p_1 must be $W \rightarrow \dots Z_k$, and we have two possibilities: (a) Z_k can be a descendent of W, and in this case the collider in U is already open even without conditioning on W, contradiction; (b) W is connected to Z_k through some collider, for instance, p_1 could be $W \rightarrow \dots \rightarrow C \leftarrow \dots Z_k$, but similarly as before, given that we condition on C, which is a descendent of W, and so of U, the collider was already conditioned as well as the path from Z_k to X open, contradiction. Therefore, it cannot be the case that after adding $\mathbf{T_k} \subseteq NonDesc(X)$ to block paths from Z_k to X or Y, there is a node Z_j such that $k < j$, and which previously had its paths to X or Y blocked, turned to have them open after conditioning on T_k. Thus, we are allowed to modify each CI obtained in step 2 before Z_k in the sequence adding $\mathbf{T_k}$, and then based on the admissible sequence starting from $OR(X, Y \mid \mathbf{C}, \mathbf{T}, Z_1, ..., Z_n)$, we can reduce it through this new augmented CIs of step 2 until reaching the desired expression $OR(X, Y \mid \mathbf{C})$.

Now we consider the complexity of the algorithm, and we show that it runs in polynomial time. Notice that only the step 3 of the algorithm could imply some backtracking – i.e., when it chooses a (minimal) set $\mathbf{T_i}$ of non-descendants of X that renders the equality in step 2 to be true. The choice of separating set *per se* is polynomial, see footnote 5.

Consider that the choice of $\mathbf{T_i}$ implies failure in step 5 when it tests the validity of $(\mathbf{T} \perp\!\!\!\perp Y \mid X, \mathbf{C})$. Assume that it exists a sequence $\mathbf{Q}$ of ancestors of S and not ancestors of X, $(Z_1, ..., Z_k, ..., Z_n)$ such that for each Z_i there is a separating set $\mathbf{T_i}$ which makes the independence test valid. Let $\mathbf{T} = \bigcup \mathbf{T_i}$, and assume that $(\mathbf{T} \perp\!\!\!\perp Y \mid X, \mathbf{C})$ holds. Assume now that in round k, the sink procedure chooses a different (minimal) separating set than $\mathbf{T_k}$, and call this new set $\mathbf{T'_k}$, and subsequently $(\mathbf{T'_{k+1}}, ..., \mathbf{T'_n})$. We have the new sequence $\mathbf{Q'}$ with additional separators $(\mathbf{T_1}, ..., \mathbf{T_{k-1}}, \mathbf{T'_k}, ..., \mathbf{T'_n})$. Call $\mathbf{T'} = \bigcup \mathbf{T'_i}$, and $\mathbf{\Delta} = \mathbf{T'} \setminus (\mathbf{T} \cap \mathbf{T'})$.

We have that $(\mathbf{T'} \not\perp\!\!\!\perp Y \mid X, \mathbf{C})$ holds, or just $(\mathbf{\Delta} \not\perp\!\!\!\perp Y \mid X, \mathbf{C})$. (This follows from $(\mathbf{\Delta} \perp\!\!\!\perp Y \mid X, \mathbf{C})$, which by composition yields $(\mathbf{T'} \perp\!\!\!\perp Y \mid X, \mathbf{C})$, contradiction. See also (Pearl and Paz, 2010).) Let $\delta \in \mathbf{\Delta}$ be the first node such that that $\mathbf{Q}$ and $\mathbf{Q'}$ disagree and which make step 5 to fail. δ blocks at least one path from Z_k to X (after condition on $\{\mathbf{C}, Y, \mathbf{T}, Z_1, ..., Z_{k-1}, \mathbf{T_i} \setminus \delta\}$) or from Z_k to Y (after condition on $\{\mathbf{C}, X, \mathbf{T}, Z_1, ..., Z_{k-1}, \mathbf{T_i} \setminus \delta\}$), otherwise the sequence will not be admissible (pass in the test of step 2). By construction, it must be the case that there is an open path from Z_k to Y passing through δ (after cond. on $\{C, X, Q, Z_1, ..., Z_{k-1}, \mathbf{T_i} \setminus \delta\}$).

Let p be part of this path from δ to Y (or, $\delta - ... - Y$).

There must exist in $\mathbf{Q}$ a vertex v which blocks this same path from Z_k to $\{X, Y\}$ or $\{Y\}$ in the test of step 2. But v is in p or connected through an open path p' to δ (i.e., $p : \delta - ... - v - ... - Y$ or $v - ... - p' - ... - \delta - ... - p - ... - Y$), otherwise we would not need δ in the first place, contradicting minimality. In both cases, there is an open path from v to Y, which contradicts the assumption about $\mathbf{Q}$ validating $(T \perp\!\!\!\perp Y \mid X, C)$ as true, and therefore it cannot exist such δ. Applying the same reasoning for the whole sequence $\mathbf{Q}'$ inductively, we conclude that it cannot exist such sequence. Therefore, step 5 does not imply any backtracking.

Similarly, let us consider the case when the choice of $\mathbf{T_j}$ implies failure in a subsequent step 2. In the sequence $\mathbf{Q}'$, it is true that when the algorithm chooses $\mathbf{T_j}$ to satisfy the admissibility of Z_j, it blocks some paths from Z_j to X. Now, assume that for Z_k, $k < j$, there is an open path through $\mathbf{T_j}$, i.e., $Z_k \longleftrightarrow U \longleftrightarrow X$, where $U = T_j$ or $T_j \in Desc(U)$. But if you do not choose $\mathbf{T_j}$ (or any other node that blocks this path), we would have an open path from Z_k to X through $\mathbf{T_j}$, contradiction.

We now argue about the completeness of the procedure. Let us first consider the case in which there is not X-independent variable in the admissible sequence, the sink-procedure will return an admissible sequence whenever one exists. Notice that the sink-procedure performs a search for an admissible sequence in reverse topological order, and this only makes the conditional independence's tests easier than in any other order. This is so because in each step, we are adding all non-descendents of Z_k (are non-colliders for Z_k), which completely disconnects Z_k from X or Y except for paths passing through non-descendents of X. (Also, non step-wise reductions can be converted to step-wise one through the graphoids decomposition and weak union.)

Assume that there is a sequence $(A_1, ..., A_m)$ called $\mathbf{A}$ that does not follow the order given by the sink-procedure and it is admissible. Now, let us call $\mathbf{Q}$ the sequence $(Z_1, ..., Z_n)$ given by the sink-procedure, and further assume that $\mathbf{Q}$ is not admissible. It is true that the last element of both sequences is S, and in $\mathbf{Q}$ we would have the blocking set $\{Z_1, ..., Z_{n-1}\}$ while in $\mathbf{A}$ we would have $\{A_1, ..., A_{m-1}\}$. It is true that $\{A_1, ..., A_{m-1}\} \subseteq \{Z_1, ..., Z_{n-1}\}$, and this is an invariant along the algorithm for all nodes in $\mathbf{A}$. Recall two facts: (a) for now, we are assuming that there are not disagreements between $\mathbf{T_Q}$ and $\mathbf{T_A}$; (b) adding descendents of Z_k in each step can only open some paths and spoil separation. It must be the case for the sink-procedure to fail, there exists $Z_k \in \mathbf{Q}$ such that $(Z_k \perp\!\!\!\perp X \mid Y, \mathbf{C}, Z_1, ..., Z_{k-1})$ and $(Z_k \perp\!\!\!\perp Y \mid X, \mathbf{C}, Z_1, ..., Z_{k-1})$ are both false. Thus, there is at least one path from Z_k to X and from Z_k to Y that are not blocked by $\{Z_1, ..., Z_{k-1}\} \cup \{C\}$ (and respectively, $\{Y\}$ and $\{X\}$); call the set of these paths P_1 and P_2, respectively.

Assume that $\mathbf{A}$ also chooses Z_k at some point along its execution, and Z_k is labeled there A_m. It must be the case that all paths from A_m to X or all paths from A_m to Y are blocked by $\{A_1, ..., A_{m-1}\} \cup \{C\}$ (and respectively, $\{Y\}$ and $\{X\}$). But if $\{A_1, ..., A_{m-1}\} \subseteq \{Z_1, ..., Z_{k-1}\}$, this is a contradiction. Now assume that $\mathbf{A}$ does not choose Z_k along its execution. There are ancestors of S which have to block P_1 from S to X or P_2 from S to Y, and we consider without loss of generality the subset $\{A_1, ..., A_l\}$ that renders this separation to hold. Consider A_j the first descendant of Z_k in G^* that is in $\{A_1, ..., A_l\}$. If such node is S, we reach a contradiction. Assume that A_j is not S but some of its ancestors. To separate A_j from X or Y, we need to block the paths from it to X or Y, but there are unblockable paths P_1 and P_2 passing through Z_k ($A_j \leftarrow ... - Z_k - P_1 - X$ or $A_j \leftarrow ... - Z_k - P_2 - Y$), and therefore A_j cannot be part of an admissible sequence, contradiction. Then, it is the case that if both algorithms do not disagree in the choice of the non-descendents of X, there is indeed not admissible sequence. For the case when we add X-independent variables along the sequence, the result also follows, and this is so based on the fact shown previously that there is no backtracking in the choice of $\mathbf{T_i}$, and any algorithm that chooses T_i consistently obtains the same outcome in terms of separation. Each time that the sink-procedure does not return any sequence, we can produce a counter-example for the G-recoverability of the triplet $(X, Y, \mathbf{C})$ based on the construction of Theorem 1. □

Theorem 4

Let us first show the result for the binary case. To match the dimensionality requirement, we assume that $\mathbf{Z} = Z_1 \cup Z_2$ and both Z_1 and Z_2 are binary satisfying:

$$P(Z_1, Z_2 \mid X, Y, S) \quad = \quad P(Z_1, Z_2 \mid X, Y) \quad (18)$$

To simplify the notation, let us write:

- $P(X = x, Y = y \mid Z_1 = z_1, Z_2 = z_2) = \alpha_{xy,z_1z_2}$
- $P(Z_1 = z_1, Z_2 = z_2) = \beta_{z_1z_2}$
- $P(Z_1 = z_1, Z_2 = z_2 \mid X = x, Y = y) = \gamma_{z_1z_2,xy}$

Note that the parameters $\gamma_{z_1z_2,xy}$ and $\beta_{z_1z_2}$ impose constraints on the distribution α_{xy,z_1z_2}, which can be made explicit by the following equation,

$$\gamma_{z_1z_2,xy} \quad = \quad \frac{\alpha_{xy,z_1z_2}\beta_{z_1z_2}}{\sum_{z_1',z_2'} \alpha_{xy,z_1'z_2'}\beta_{z_1'z_2'}} \quad (19)$$

M	1	2	3	4	5	6	7	8	9	10	11	12
1	$(c_1-1)b_1$	c_1b_2	c_1b_3	c_1b_4								
2	c_2b_1	$(c_2-1)b_2$	c_2b_3	c_2b_4								
3	c_3b_1	c_3b_2	$(c_3-1)b_3$	c_3b_4								
4					$(c_4-1)b_1$	c_4b_2	c_4b_3	c_4b_4				
5					c_5b_1	$(c_5-1)b_2$	c_5b_3	c_5b_4				
6					c_6b_1	c_6b_2	$(c_6-1)b_3$	c_6b_4				
7									$(c_7-1)b_1$	c_7b_2	c_7b_3	c_7b_4
8									c_8b_1	$(c_8-1)b_2$	c_8b_3	c_8b_4
9									c_9b_1	c_9b_2	$(c_9-1)b_3$	c_9b_4
10	$(1-c_{10})b_1$	$-c_{10}b_2$	$-c_{10}b_3$	$-c_{10}b_4$	$(1-c_{10})b_1$	$-c_{10}b_2$	$-c_{10}b_3$	$-c_{10}b_4$	$(1-c_{10})b_1$	$-c_{10}b_2$	$-c_{10}b_3$	$-c_{10}b_4$
11	$-c_{11}b_1$	$(1-c_{11})b_2$	$-c_{11}b_3$	$-c_{11}b_4$	$-c_{11}b_1$	$(1-c_{11})b_2$	$-c_{11}b_3$	$-c_{11}b_4$	$-c_{11}b_1$	$(1-c_{11})b_2$	$-c_{11}b_3$	$-c_{11}b_4$
12	$-c_{12}b_1$	$-c_{12}b_2$	$(1-c_{12})b_3$	$-c_{12}b_4$	$-c_{12}b_1$	$-c_{12}b_2$	$(1-c_{12})b_3$	$-c_{12}b_4$	$-c_{12}b_1$	$-c_{12}b_2$	$(1-c_{12})b_3$	$-c_{12}b_4$

Now, for a given assignment $< X=0, Y=0 >$, let us list all independent parameters $\gamma_{z_1z_2,00}$,

$$\begin{aligned}\gamma_{00,00} &= \frac{\alpha_{00,00}\beta_{00}}{\sum_{z_1',z_2'}\alpha_{00,z_1'z_2'}\beta_{z_1'z_2'}}\\ \gamma_{01,00} &= \frac{\alpha_{00,01}\beta_{01}}{\sum_{z_1',z_2'}\alpha_{00,z_1'z_2'}\beta_{z_1'z_2'}}\\ \gamma_{10,00} &= \frac{\alpha_{00,10}\beta_{10}}{\sum_{z_1',z_2'}\alpha_{00,z_1'z_2'}\beta_{z_1'z_2'}}\end{aligned} \tag{20}$$

Note that $\gamma_{11,00}$ is not an independent parameter because it is completely determined by the other three equations in (20) given the integrality constraint. For now, we have 3 equations and 4 unknown variables ($\{\alpha_{00,00}, \alpha_{00,01}, \alpha_{00,10}, \alpha_{00,11}\}$.)

Similarly, we write the constraints for the assignments $< X=1, Y=0 >$ and $< X=0, Y=1 >$, respectively,

$$\gamma_{00,10} = \frac{\alpha_{10,00}\beta_{00}}{\sum_{z_1',z_2'}\alpha_{10,z_1'z_2'}\beta_{z_1'z_2'}}, ... \tag{21}$$

$$\gamma_{00,01} = \frac{\alpha_{01,00}\beta_{00}}{\sum_{z_1',z_2'}\alpha_{01,z_1'z_2'}\beta_{z_1'z_2'}}, ... \tag{22}$$

Now, we can write the equations for the constraints relative to the variables α_{11,z_1z_2} as a function of the previous variables $\{\alpha_{00,z_1z_2}, \alpha_{01,z_1z_2}, \alpha_{10,z_1z_2}\}$,

$$\gamma_{00,11} = \Big((1-\big(\alpha_{00,00}+\alpha_{01,00}+\alpha_{10,00}\big))\beta_{00}\Big)\Big/ \Big(\sum_{Z_1',Z_2'}\Big(1-\big(\alpha_{00,z_1'z_2'}+\alpha_{01,z_1'z_2'}+\alpha_{10,z_1'z_2'}\big)\Big)\beta_{z_1'z_2'}\Big), ... \tag{23}$$

Notice that the parameters $\gamma_{z_1z_2,11}$ are independent, and we have 12 equations and 12 unknowns, but it remains to show that the equations are all independent (notice that the last three constraints in eq. (23) involve variables of the other constraints). Another fact to observe is that the system is indeed linear. We show that the matrix M, induced by the eqs. (20, 21, 22, 23), is linear and (almost surely) invertible, and generates an unique solution. M is invertible if and only if its determinant is non-zero. For convenience, let us display the variables α_{xy,z_1z_2} column-wise, renaming $\beta_{z_1z_2}$ as constants b_1-b_4, and $\gamma_{z_1z_2,xy}$ as constants c_1-c_{12}. The matrix is shown on the top of the previous page.

In what follows, we exploit the block structure of M and apply the following transformations to better visualize its determinant.

1. First note that all columns $\{1,5,9\}$ are multiplied by b_1, which can be factored out by the determinant property. Similarly for the other columns in respect to $\{b_2,b_3,b_4\}$, which can be expressed as $det(M) = (b_1b_2b_3b_4)^3 det(M^{(1)})$, where $M^{(1)}$ is the resultant matrix.
2. Let us sum lines $\{1,4,7\}$ to line 10, lines $\{2,5,8\}$ to line 11, and $\{3,6,9\}$ to line 12, which generate matrix $M^{(2)}$.
3. We now sum the columns of $M^{(2)}$, -1 times column 4 to column 1, -1 times column 4 to column 2, and -1 times column 4 to column 3 (similarly for the other blocks), which yields $M^{(3)}$.
4. Sum the columns of $M^{(3)}$, c_1 times column 1, c_2 times column 2 and c_3 times column 3 to column 4 (similarly for the other blocks), yielding $M^{(4)}$.
5. Now, reorder the columns, "pushing" column 4 and 8 towards the end, call the resultant matrix $M^{(5)}$.

Now we are done, notice that the $det(M) = (b_1b_2b_3b_4)^3 det(M^{(5)})$, and the determinant of $M^{(5)}$ is the determinant of two block matrices, the square matrix $M_1^{(5)}$ from lines 1-9 multiplied by another square matrix $M_2^{(5)}$ from lines 10-12. Note that $det(M_1^{(5)}) = -1$, and remains to show that $det(M_2^{(5)})$ is almost always different than zero. The parameters c_1 to c_{12} are independent, and given the form obtained to $M_2^{(5)}$ where all entries are independent, this implies that $M_2^{(5)}$ is non-singular almost surely, and so it is $M^{(5)}$ – coincidental cancellations will occur with Lebesgue measure zero.

Therefore, we consider M as full rank, which can be solved algebraically with standard techniques yielding the solution $\alpha = M^{-1}\gamma$. This result, together with $P(\mathbf{Z})$ yields the joint distribution $P(Y,X,\mathbf{Z})$. The case for non-binary variables follows in a straightforward way, just noticing the requirement for agreement between the dimensions of the IV set $\mathbf{Z}$ and $\{X,Y\}$. □

Corollary 5

First, apply Theorem 4 to the variables $\{W, Y\}$ replacing X with W, and obtain $P(W, Y)$. Further note that $P(X \mid Y, W, S = 1) = P(X \mid Y, W)$, which together with the first observation finishes this part of proof. The proof for when we do not rely on $\mathbf{Z}$ is essentially the same. □

References

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92** 1172–1176.

Cooper, G. (1995). Causal discovery from data in the presence of selection bias. *Artificial Intelligence and Statistics* 140–150.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute* **11** 1269–1275.

Didelez, V., Kreiner, S. and Keiding, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statistical Science* **25(3)** 368–387.

Evans, R. J. and Richardson, T. S. (2011). Marginal log-linear parameters for graphical markov models. arXiv:1105.6075 [stat.ME].

Geneletti, S., Richardson, S. and Best, N. (2009). Adjusting for selection bias in retrospective, case-control studies. *Biostatistics* **10(1)**.

Geng, Z. (1992). Collapsibility of relative risk in contingency tables with a response variable. *Journal Royal Statistical Society* **54** 585–593.

Glymour, M. and Greenland, S. (2008). Causal diagrams. In *Modern Epidemiology* (K. Rothman, S. Greenland and T. Lash, eds.), 3rd ed. Lippincott Williams & Wilkins, Philadelphia, PA, 183–209.

Greenland, S. and Pearl, J. (2011). Adjustments and their consequences – collapsibility analysis using graphical models. *International Statistical Review* **79** 401–426.

Heckman, J. (1970). Sample selection bias as a specification error. *Econometrica* **47** 153–161.

Hein, M. (2009). Binary classification under sample selection bias. In *Dataset Shift in Machine Learning* (J. Candela, M. Sugiyama, A. Schwaighofer and N. Lawrence, eds.). MIT Press, Cambridge, MA, 41–64.

Hernán, M., Hernández-Díaz, S. and Robins, J. (2004). A structural approach to selection bias. *Epidemiology* **15** 615–625.

Horwitz, R. and Feinstein, A. (1978). Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *New England Journal of Medicine* **299** 368–387.

Lauritzen, S. L. and Richardson, T. S. (2008). Discussion of mccullagh: Sampling bias and logistic models. *J. Roy. Statist. Soc. Ser. B* **70** 140–150.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 425–432.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 247–254.

Pearl, J. and Paz, A. (2010). Confounding equivalence in causal equivalence. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 433–441.

Robins, J. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **12** 313–320.

Robins, J. M., Hernan, M. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

Smith, A. T. and Elkan, C. (2007). Making generative classifiers robust to selection bias. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '07, ACM, New York, NY, USA.

Storkey, A. (2009). When training and test sets are different: characterising learning transfer. In *Dataset Shift in Machine Learning* (J. Candela, M. Sugiyama, A. Schwaighofer and N. Lawrence, eds.). MIT Press, Cambridge, MA, 3–28.

Tian, J., Paz, A. and Pearl, J. (1998). Finding minimal separating sets. Tech. Rep. R-254, <http://ftp.cs.ucla.edu/pub/stat_ser/r254.pdf>, Computer Science Department, University of California, Los Angeles, CA.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence, Proceedings of the Sixth Conference*. Cambridge, MA. Also in P. Bonissone, M.

Henrion, L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 6*, Elsevier Science Publishers, B.V., 255–268, 1991.

Whittemore, A. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society, Series B* **40** 328–340.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *ICML* (C. E. Brodley, ed.), vol. 69 of *ACM International Conference Proceeding Series*. ACM.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **172** 1873–1896.